Distributed Random Projection Algorithm for Convex Optimization

Soomin Lee and Angelia Nedić

Abstract

Random projection algorithm is of interest for constrained optimization when the constraint set is not known in advance or the projection operation on the whole constraint set is computationally prohibitive. This paper presents a *distributed random projection* (DRP) algorithm for fully distributed constrained convex optimization problems that can be used by multiple agents connected over a time-varying network, where each agent has its own objective function and its own constrained set. With reasonable assumptions, we prove that the iterates of all agents converge to the same point in the optimal set almost surely. In addition, we consider a variant of the method that uses a mini-batch of consecutive random projections and establish its convergence in almost sure sense. Experiments on distributed support vector machines demonstrate fast convergence of the algorithm. It actually shows that the number of iteration required until convergence is much smaller than scanning over all training samples just once.

I. INTRODUCTION

A number of problems that arise in sensor, wireless *ad hoc* and peer-to-peer networks can be formulated as convex constrained minimization problems [1]–[4]. The goal of the agents connected over such networks is to cooperatively solve the following optimization problem:

$$\min_x f(x) = \sum_{i=1}^{m} f_i(x) \qquad \text{s.t. } x \in \mathcal{X} \triangleq \bigcap_{i=1}^{m} \mathcal{X}_i, \tag{1}$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex function, representing the local objective function of agent i, and each $\mathcal{X}_i \subseteq \mathbb{R}^d$ is a closed convex set, representing the local constraint set of agent i. The complete problem information is not available at a single location. This is because i) there is no central node that facilitates computation and communication and ii) it is often not possible for one agent to keep all the objective and constraint components due to memory, computational power, or privacy constraints. In addition, the network topology itself may change with time due to agent mobility or link failures. Therefore, an optimization algorithm for solving such problems must be distributed and robust, so that each agent exchanges its information only with its immediate neighbors and the algorithm has to be adaptive to the changes in the network topology.

In this paper, we propose a *distributed random projection* (DRP) algorithm for problem (1), where the constraint set is defined as the intersection of finitely many simple convex constraints. That is, $\mathcal{X}_i = \bigcap_{j \in I_i} \mathcal{X}_i^j$, where I_i is a finite[1] (a formal definition of I_i is in Section II). In our algorithm, each agent i maintains its own iterate sequence $\{x_i(k)\}$. At each iteration, each agent calculates weighted average of the received iterates (from its neighbors) and its own iterate, adjusts the iterate by using gradient information of its local objective function f_i and projects onto a constraint component that is selected randomly from its local constraint set $\mathcal{X}_i$. The projections are performed locally by each agent based on the random observations of the local constraint components. In particular, agent i observes a constraint component $\mathcal{X}_i^{\Omega_i(k)}$ at time k, where $\Omega_i(k) \in I_i$ is a random variable.

Our primary interest is in the case when the whole constraint set $\mathcal{X}_i$ for an agent i is not known in advance, but its component is revealed through random realizations $\mathcal{X}_i^{\Omega_i(k)}$. For example, in collaborative filtering for recommender systems, user data is huge and distributed over multiple machines. Users frequently change and update their preferences in real time so the constraint set of this problem is usually not known in advance. Another case of interest

[1]The finiteness of I_i is not really crucial. The developed results also apply to the case when the index sets I_i are infinite.

is when the whole constraint set $\mathcal{X}_i$ is known in advance but it has a huge number of components. For example, in text classification problems, model parameters are trained based on hundred thousands or more text samples and each sample constitutes a constraint component (usually a halfspace) [5]. In such a case, the projection operation on the whole constraint set $\mathcal{X}_i$ is computationally prohibitive if any of the traditional (sub)gradient projection methods are used. In Section VII, we will experiment with Support Vector Machines to classify three text data sets.

In the optimization literature, algorithms of two categories have been proposed for problem (1): the Markov incremental algorithm and the distributed subgradient algorithm. In the Markov incremental algorithm studied in [6], [7], the agents maintain a single estimate sequence that is sequentially updated by one agent at a time. When an agent receives the estimate, it updates the estimate using its local objective function and passes it to a randomly selected neighbor. The update order is driven by a time inhomogeneous Markov chain (as the network topology is time varying). Whereas in the distributed subgradient algorithms, each agent maintains its own estimate. It communicates the estimate with its neighbors and updates it using the local objective and constraint information. Algorithms of this type requires a consensus over all agents for convergence. However, in some distributed problems it is important that each agent maintains a good estimate at all times. For example, in a distributed online learning, each node is expected to perform in real time. Our DRP algorithm is in the distributed subgradient algorithm category.

The related distributed optimization literature includes [8]–[15], which are concerned with convex but unconstrained problems, and [16]–[18] where constrained problems are considered. The most relevant to the work in this paper are [19]–[22] where, as in the DRP algorithm, the constraint set is also distributed across agents and each agent handles its own constraint set only. In [19], the convergence analysis is done for a special case when the network is completely connected. The work in [20], [21] extends the algorithm and its analysis to a more general network including the presence of noisy links, while [22] extends it to a general Markovian network model. Unlike [19] and [20], where each agent can perform projections on its entire constraint set, this paper addresses the case when such projections are not possible or computationally prohibitive. Related to this work are also the distributed algorithms for estimation and inference problems that have been proposed and studied by Sayed et al. [23]–[26]. On a much broader scale, the work in this paper is related to the literature on the consensus problem, where each agent starts from an initial value and ends by converging to a value common to all agents (see for example [4], [27]–[30]).

The contribution of this paper is mainly in two directions. First, we propose a novel distributed optimization algorithm that is based on local communications of agents' estimates in a network and a gradient descent with random projections. Second, we study the convergence of the algorithm and its variant using a mini-batch of random projections. To the best of our knowledge, there is no previous work on distributed optimization algorithms that utilize random projections. Gradient and subgradient random projection algorithms for *centralized* (not distributed) convex problems have been proposed in [31]. Also, the related work is the (centralized) random projection method for a special class of convex feasibility problems, which has been proposed and studied by Polyak [32].

The rest of the paper is organized as follows. In Section II, we introduce the problem of interest, formally describe our algorithm and state assumptions on the problem and network. In Section III, we state some results from the literature that we use in the convergence analysis. In Section IV, we derive two important results that will play crucial roles in the convergence analysis. In Section V, we study the almost sure convergence property of our DRP algorithm. We provide an extension of the algorithm to a variant that uses a mini-batch of random projections and we state a convergence result for this extension in Section VI. As an application of our DRP algorithm and its mini-batch variant, in Section VII, we introduce a linear SVM formulation, discuss how to apply the algorithm, and present some experimental results on binary text classification tasks. Section VIII contains concluding remarks and future directions.

Notation A vector is viewed as a column. We write x^T to denote the transpose of a vector x. The scalar product of two vectors x and y is $\langle x, y \rangle$. We use a subscript i to denote an agent i. An index k with parentheses is devoted to represent a time. For example, $x_i(k)$ is the iterate of an agent i at time k. We use $\|x\|$ to denote the standard Euclidean norm. We write $\text{dist}(x, \mathcal{X})$ for the distance of a vector x from a closed convex set $\mathcal{X}$, *i.e.*, $\text{dist}(x, \mathcal{X}) = \min_{v \in \mathcal{X}} \|v - x\|$. We use $\Pi_{\mathcal{X}}[x]$ for the projection of a vector x on the set $\mathcal{X}$, *i.e.*, $\Pi_{\mathcal{X}}[x] = \arg\min_{v \in \mathcal{X}} \|v - x\|^2$. We use $\Pr\{Z\}$ and $\mathsf{E}[Z]$ to denote the probability and the expectation of a random variable Z. We abbreviate *almost surely* and *independent and identically distributed* as *a.s.* and *iid*, respectively.

II. PROBLEM SET-UP, ALGORITHM AND ASSUMPTIONS

A. *Optimization over a Network*

We consider a constrained convex optimization problem (1) that is distributed over a network of m agents, indexed by $V = \{1, \ldots, m\}$. The function f_i and the constraint set $\mathcal{X}_i$ in (1) are private information of agent i (not shared with any other agent). Collectively, the agents are responsible for solving problem (1).

We are interested in the case when each constraint set $\mathcal{X}_i$ is the intersection of finitely many closed convex sets. Without loss of generality, let $\mathcal{X}$ be the intersection of n closed convex sets. Let $I = \{1, \ldots, n\}$ be the index set, and let I_i, $i \in V$, be a partition of I (i.e., $I = \bigcup_{i=1}^{m} I_i$ and $I_i \cap I_j = \emptyset$ for $i \neq j$) such that each I_i is associated with the local constraint set $\mathcal{X}_i$ of agent i, i.e.,

$$\mathcal{X}_i = \cap_{j \in I_i} \mathcal{X}_i^j \qquad \text{for a finite index set } I_i,$$

where the superscript is used to identify a component set. Each component set $\mathcal{X}_i^j$ is assumed to be a "simple set" for the projection operation. Examples of such simple sets include a halfspace $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \langle a, x \rangle \leq b\}$, a box $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \alpha \leq x \leq \beta\}$ (the inequality is component-wise) and a ball $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \|x - v\| \leq r\}$, where $a, \alpha, \beta, v \in \mathbb{R}^d$ and $b, r \in \mathbb{R}$. In such cases, the projection on the set $\mathcal{X}_i$ can be complex, especially when the number of components is large, while the projection on each component $\mathcal{X}_i^j$ has a closed form expression.

We use the following assumption for the functions f_i and the sets $\mathcal{X}_i^j$.

Assumption 1: Let the following conditions hold:

(a) The sets $\mathcal{X}_i^j$, $j \in I_i$ are closed and convex for every $i \in V$.

(b) Each function $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex.

(c) The functions f_i, $i \in V$, are differentiable and have *Lipschitz gradients* with a constant L over $\mathbb{R}^d$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

(d) The gradients $\nabla f_i(x)$, $i \in V$ are bounded over the set $\mathcal{X}$, i.e., there exists a constant G_f such that

$$\|\nabla f_i(x)\| \leq G_f \quad \text{for all } x \in \mathcal{X} \text{ and all } i \in V.$$

When each f_i has Lipschitz gradients with a constant L_i, Assumption 1(c) is satisfied with $L = \max_{i \in V} L_i$. Further note that Assumption 1(d) is satisfied, for example, when $\mathcal{X}$ is compact.

As mentioned earlier, the agents are collectively responsible for solving problem (1), without sharing their private knowledge of individual objective functions f_i and the constrained sets $\mathcal{X}_i$. To accommodate such a task, the agents are assumed to form a network, wherein each agent communicates its iterates to its local neighbors. More specifically, at each time k, the network topology is represented by a directed graph $G(k) = (V, E(k))$, where $E(k) \subseteq V \times V$. A link $(i, j) \in E(k)$ indicates that agent i has received information from agent j at time k. We let $N_i(k)$ denote the set of agents who send information to agent i, i.e., $N_i(k) = \{j \in V \mid (i, j) \in E(k)\}$. We assume that $i \in N_i(k)$ for all $i \in V$ and for all k.

B. *Distributed Random Projection Algorithm (DRP)*

To solve the problem (1) with distributed information access, we propose an iterative gradient method with random projections. Let $x_i(k) \in \mathbb{R}^d$ denote the estimate of agent i at time k. At time k, each agent sends the estimate to its neighbors (represented by the graph $(V, E(k))$). Upon receiving the estimates $x_j(k)$ from its neighbors $j \in N_i(k)$, each agent i updates according to the following two steps:

$$v_i(k) = \sum_{j \in N_i(k)} w_{ij}(k)x_j(k) \tag{2a}$$

$$x_i(k+1) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}} \left[v_i(k) - \alpha_k \nabla f_i(v_i(k))\right], \tag{2b}$$

where $\alpha_k > 0$ is a stepsize at time k and $x_i(0) \in \mathbb{R}^d$ is an initial estimate of agent i (which can be random).

In the above, relation (2a) captures an information mixing step, while (2b) captures a local minimization and feasibility update step using a random projection. In (2a), the iterate $v_i(k)$ is a weighted average of agent i's estimate and the estimates received from its neighbors $j \in N_i(k)$. Specifically, $w_{ij}(k) \geq 0$ is a weight that agent i

places on the estimate $x_j(k)$ received from a neighbor $j \in N_i(k)$ at time k, where the total weight sum is 1, i.e., $\sum_{j \in N_i(k)} w_{ij}(k) = 1$ for each agent i. The step (2a) can be equivalently represented as

$$v_i(k) = \sum_{j=1}^{m} [W(k)]_{ij} x_j(k) \tag{3}$$

by letting $w_{ij}(k) = 0$ for whenever $j \notin N_i(k)$, and using $[W]_{ij}$ to denote the (i, j)th entry of a matrix W.

In (2b), agent i adjusts the average $v_i(k)$ along the negative gradient direction of its local objective f_i. At time k, agent i also observes a random realization of its local constraint component set $\mathcal{X}_i^{\Omega_i(k)}$. To reduce the feasibility violation, it projects its current estimate on this set. The random variable $\Omega_i(k)$ takes values in the index set I_i at all times k. In this way, instead of projecting onto the whole local constraint set $\mathcal{X}_i$, agent i projects only on a component set $\mathcal{X}_i^{\Omega_i(k)}$ which is randomly selected at time k. Note that the updated estimate $x_i(k+1)$ may not lie in $\mathcal{X}_i$ since $\mathcal{X}_i \subset \mathcal{X}_i^{\Omega_i(k)}$.

Through the updates (2a) and (2b), agents combine their information and consider their own optimization problem of minimizing f_i over the set $\mathcal{X}_i$. There is neither a central node governing the whole process nor additional constraints enforcing consistency. Nevertheless, with this simple update rule, our algorithm finds the optimal solution and all agents eventually arrive at a common optimal solution (all $x_i(k)$ converge to some $x^* \in \mathcal{X}^*$, as shown in Section V).

Note that algorithm (2a)-(2b) is similar to the distributed projected subgradient algorithm in [19] except for the randomization over the components of the set $\mathcal{X}_i$ in (2b). At each iteration of the algorithm in [19], a projection is performed on the entire constraint set $\mathcal{X}_i$, which can be prohibitively expensive when $\mathcal{X}_i$ is itself an intersection of many sets. In addition, unlike the method in [19], DRP can also handle the cases when the projection on the entire set $\mathcal{X}_i$ is not possible since the set $\mathcal{X}_i$ may not be known in advance.

The challenges in convergence analysis of the DRP algorithm are posed mainly by its distributed nature, through the *effects of the time-varying network*, and by the *projection errors* associated with using projections on components $\mathcal{X}_i^j$, $j \in I_i$ of the set $\mathcal{X}_i = \cap_{j \in I_i} \mathcal{X}_i^j$ instead of the projection on the set $\mathcal{X}_i$. The fact that the DRP relies on a random component $\mathcal{X}_i^j$ poses particular difficulties, as one needs to characterize the impact of the random projection errors, which is closely related to errors in "set-approximations". To handle these difficulties, we make several mild assumptions. We make an assumption on the random set processes $\{\Omega_i(k)\}$, $i \in V$, that allows us to characterize the projection errors. For the network we assume that it is sufficiently connected in order to properly conduct the information among the agents. Finally, we assume that the agent weights are also properly chosen to ensure that each agent is equally influencing every other agent. These network assumptions have been typically used in distributed optimization algorithms over a time-varying network (see e.g. [13], [14], [16], [33]–[35]). In the next subsections, we state our assumptions on the random set processes $\{\Omega_i(k)\}$, $i \in V$, the network and the weight matrices $W(k)$.

C. Assumptions on Random Set Process

For the random sequences $\{\Omega_i(k)\}$, $i \in V$, we assume the following.

Assumption 2: The sequences $\{\Omega_i(k)\}$, $i \in V$, are iid and independent of the initial random points $x_i(0)$, $i \in V$. We have $\pi_i^j \triangleq \Pr\{\Omega_i(k) = j\} > 0$ for all $j \in I_i$ and $i \in V$.

The variable $\Omega_i(k)$ can be viewed as a random sample at time k of a random variable Ω_i that takes values $j \in I_i$ with probability π_i^j. In some situations the probability distributions π_i may be dictated by nature and agent i cannot control them. In situations where the agents have all sets $\mathcal{X}_i^j$, $j \in I_i$ available, each agent i can choose a uniform distribution π_i over the set I_i.

The next assumption is crucial in our analysis.

Assumption 3: For all $i \in V$, there exists a constant $c > 0$ such that for all $x \in \mathbb{R}^d$,

$$\text{dist}^2(x, \mathcal{X}) \leq c\mathsf{E}\left[\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)})\right]. \tag{4}$$

Assumption 3 is satisfied, for example, when each set $\mathcal{X}_i^j$ is given by either linear inequality or a linear equality, or when the intersection set $\mathcal{X}$ has a nonempty interior. In the first case, one can verify that the assumption holds by using the results of Burke and Ferris on a set of weak sharp minima [36]. In the second case, one can use the

ideas of the convergence rate analysis for the alternating projection algorithm of Gubin, Polyak and Raik in [37]. In either case, the constant c depends on the probability distributions π_i and some geometric properties of the sets.

D. Assumptions on the Network and Weight Matrices

We rely on the graphs $(V, E(k))$, $k \geq 0$ to represent the time-varying network. We make two assumptions.

Assumption 4: [Network Connectivity] There exists a scalar Q such that the graph $\left(V, \bigcup_{\ell=0,\ldots,Q-1} E(k+\ell)\right)$ is strongly connected for all $k \geq 0$.

Assumption 4 ensures that the agents communicate sufficiently often so that all functions and all constraints (f_i's and $\mathcal{X}_i$'s) influence the iterates of all agents.

Next, we make the following assumption on the edge weights (defined below (3)).

Assumption 5: [Doubly Stochasticity] For all $k \geq 0$,
(a) $[W(k)]_{ij} \geq 0$ and $[W(k)]_{ij} = 0$ when $j \notin N_i(k)$,
(b) $\sum_{j=1}^{m} W[(k)]_{ij} = 1$ for all $i \in V$,
(c) There exists a scalar $\eta \in (0,1)$ such that $[W(k)]_{ij} \geq \eta$ when $j \in N_i(k)$,
(d) $\sum_{i=1}^{m} [W(k)]_{ij} = 1$ for all $j \in V$.

Assumption 5(a) states that the weights respect the network topology at any time k. Assumption 5(b) means that each agent calculates a weighted average of the estimates obtained from its neighbors. Assumption 5(c) ensures that each agent gives sufficient weights on the information received. Assumption 5(d) together with Assumption 4 ensure that each agent is equally influential in the long run so that the agents arrive at a consensus on an optimal solution.

III. PRELIMINARIES

In this section, we state some definitions and results from the literature, which will be used in later sections.

Convexity of Euclidean norm and its square. Both the Euclidean norm and its square are convex functions, i.e., for any vectors $v_1, \ldots, v_m \in \mathbb{R}^d$ and nonnegative scalars $\beta_1, \ldots, \beta_m$ such that $\sum_{i=1}^{m} \beta_i = 1$, we have

$$\left\| \sum_{i=1}^{m} \beta_i v_i \right\| \leq \sum_{i=1}^{m} \beta_i \|v_i\|, \quad \left\| \sum_{i=1}^{m} \beta_i v_i \right\|^2 \leq \sum_{i=1}^{m} \beta_i \|v_i\|^2. \tag{5}$$

Non-expansive projection property. We state a projection theorem (see [38] for its proof).

Lemma 1: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. The function $\Pi_{\mathcal{X}} : \mathbb{R}^d \to \mathcal{X}$ is continuous and nonexpansive, i.e.,
(a) $\|\Pi_{\mathcal{X}}[x] - \Pi_{\mathcal{X}}[y]\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^d$.
(b) $\|\Pi_{\mathcal{X}}[x] - y\|^2 \leq \|x - y\|^2 - \|\Pi_{\mathcal{X}}[x] - x\|^2$ for all $x \in \mathbb{R}^d$ and for all $y \in \mathcal{X}$.

Matrix convergence. Recall we defined $W(k)$ to be the matrix with (i,j)th entry equal to $w_{ij}(k)$. From Assumption 5, the matrix $W(k)$ is doubly stochastic. Define for all k, s with $k > s \geq 0$,

$$\Phi(k, s) = W(k)W(k-1) \cdots W(s+1)W(s), \tag{6}$$

with $\Phi(k, k) = W(k)$ for all $k \geq 0$. We state the convergence property of the matrix $\Phi(k, s)$ (see [14] for its proof). Let $[\Phi(k,s)]_{ij}$ denote the (i,j)th entry of the matrix $\Phi(k, s)$, and $e \in \mathbb{R}^m$ be the column vector whose all entries are equal to 1.

Lemma 2: Let Assumptions 4 and 5 hold. Then,
(a) $\lim_{k \to \infty} \Phi(k, s) = \frac{1}{m} e e^T$ for all $s \geq 0$.
(b) $\left| [\Phi(k,s)]_{ij} - \frac{1}{m} \right| \leq \theta \beta^{k-s}$ for all $k \geq s \geq 0$, where $\theta = \left(1 - \frac{\eta}{4m^2}\right)^{-2}$ and $\beta = \left(1 - \frac{\eta}{4m^2}\right)^{\frac{1}{Q}}$.

Supermartingale convergence result. In our analysis of the DRP algorithm, we also make use of the following supermartingale convergence result due to Robbins and Siegmund (see [39, Lemma 10-11, p. 49-50]).

Theorem 1: Let $\{v_k\}$, $\{u_k\}$, $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative random variables such that

$$\mathsf{E}[v_{k+1} | \mathcal{F}_k] \leq (1 + a_k) v_k - u_k + b_k \quad \text{for all } k \geq 0 \quad a.s.,$$

where $\mathcal{F}_k$ denotes the collection $v_0, \ldots, v_k$, $u_0, \ldots, u_k$, $a_0, \ldots, a_k$ and $b_0, \ldots, b_k$. Also, let $\sum_{k=0}^{\infty} a_k < \infty$ and $\sum_{k=0}^{\infty} b_k < \infty$ a.s. Then, we have $\lim_{k \to \infty} v_k = v$ for a random variable $v \geq 0$ *a.s.*, and $\sum_{k=0}^{\infty} u_k < \infty$ *a.s.*

The above theorem is the key in our convergence analysis. Specifically, once we show that Theorem 1 applies to $v_{k+1} = \sum_{i=1}^{m} \|x_i(k+1) - x^*\|^2$ for an optimal solution x^*, the rest of the proof just builds on the implications of the theorem.

Scalar Sequences. We also use the convergence result for scalar sequences (see Lemma 3.1 in [16] for its proof). For a scalar β and a scalar sequence $\{\gamma(k)\}$, we consider the convolution sequence $\sum_{\ell=0}^{k} \beta^{k-\ell} \gamma(\ell)$.

Lemma 3: If $\lim_{k \to \infty} \gamma(k) = \gamma$ and $0 < \beta < 1$, then $\lim_{k \to \infty} \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma(\ell) = \frac{\gamma}{1-\beta}$.

IV. BASIC RELATIONS

Our convergence analysis is based on a critical relation that captures the decrease in values $\sum_{i=1}^{m} \|x_i(k+1) - x^*\|^2$ as the algorithm progresses. Such a relation is provided in Lemma 4, which is taken from [31] where it was developed for a centralized algorithm. This basic relation is further refined to take into account the distributed nature of the algorithm. Specifically, in Lemma 5, we show that the weighted averages $v_i(k)$ of the iterates approach the constraint set X asymptotically. Then, in Lemma 7, we prove that the agents' disagreement on $v_i(k)$ is diminishing with the number k of iterations. The proof of Lemma 7 relies on an auxiliary result taken from [16], which is provided in Lemma 6.

In the analysis, we will rely on the expectation taken with respect to the past history of the algorithm, which we define as follows. Let $\mathcal{F}_k$ be the σ-algebra generated by the entire history of the algorithm up to time $k-1$ inclusively (realizations of all the random variables but not the realizations of the indices Ω_i at time k), i.e., for all $k \geq 1$,

$$\mathcal{F}_k = \{x_i(0), i \in V\} \cup \{\Omega_i(\ell); 0 \leq \ell \leq k-1, i \in V\},$$

where $\mathcal{F}_0 = \{x_i(0), i \in V\}$. Therefore, given $\mathcal{F}_k$, the collection $x_i(0), \ldots, x_i(k)$ and $v_i(0), \ldots, v_i(k)$ generated by the algorithm (2a)-(2b) is fully determined.

A. Basic Iterate Relation

The following lemma is from the paper [31, Lemma 1], which provides relation among the iterate obtained after one step of the algorithm (2a), a point in the feasible set $\mathcal{X}$ and an arbitrary point in $\mathbb{R}^d$.

Lemma 4: Let $\mathcal{Y}$ be a closed convex set such that $\mathcal{Y} \subseteq \mathbb{R}^d$. Let the function $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable over $\mathbb{R}^d$ with Lipschitz continuous gradients with a constant L. Let y be given by

$$y = \Pi_{\mathcal{Y}}[x - \alpha \nabla \phi(x)] \quad \text{for some } x \in \mathbb{R}^d \text{ and } \alpha > 0.$$

Then, we have for any $\check{x} \in \mathcal{Y}$ and $z \in \mathbb{R}^d$,

$$\begin{aligned}
\|y - \check{x}\|^2 \leq &(1 + A_\tau \alpha^2)\|x - \check{x}\|^2 - 2\alpha(\phi(z) - \phi(\check{x})) \\
&- \frac{3}{4}\|y - x\|^2 + \left(\frac{3}{8\tau} + 2\alpha L\right)\|x - z\|^2 \\
&+ B_\tau \alpha^2 \|\nabla \phi(\check{x})\|^2,
\end{aligned} \tag{7}$$

where $A_\tau = 8L^2 + 16\tau L^2$, $B_\tau = 8\tau + 8$ and $\tau > 0$ is arbitrary.

Lemma 4 provides a measure of progress toward an optimal point of the function ϕ when moving from a point x in the direction opposite of the gradient $\nabla \phi(x)$. Specifically, if x^* is a minimizer of $\phi(x)$ over $\mathcal{Y}$, the lemma (with $\check{x} = x^*$) will provide us with a relation between the distances $\|y - x^*\|$ and $\|x - x^*\|$, where the point y is resulting from a projected-gradient step away from the point x. The lemma provides a relation that helps us measure the progress of a gradient-based algorithm for minimizing ϕ. Lemma 4, with a specific identification of the terms, will be a starting point for our convergence proof.

B. Projection Estimate

In the next lemma, we show that the sequences $\{v_i(k)\}$, $i \in V$, approach the constraint set $\mathcal{X}$. The result does not say that these sequences necessarily have accumulation points in $\mathcal{X}$, but rather that the distance between $v_i(k)$

and the set $\mathcal{X}$ tends to 0, as $k \to \infty$, for all i. Furthermore, these distances converge to 0 rather fast, as the sum of all squared distances over time is finite, which is a critical relation in our analysis.

Lemma 5: Let Assumption 1 hold. Let each $W(k)$ be doubly stochastic, and let $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Then,

$$\sum_{k=0}^{\infty} \text{dist}^2(v_i(k), \mathcal{X}) < \infty \qquad \text{for all } i \in V \ a.s.$$

Proof: In Lemma 4, let $y = x_i(k+1)$, $x = v_i(k)$, $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$, $\alpha = \alpha_k$, $\phi = f_i$ and $\tau = c$ where c is the constant from Assumption 3. Then, for any $\check{x} \in \mathcal{X}$ (also in $\mathcal{X}_i^{\Omega_i(k)}$, since $\mathcal{X} \subseteq \mathcal{X}_i^{\Omega_i(k)}$) and any $z \in \mathbb{R}^d$, we obtain

$$\|x_i(k+1) - \check{x}\|^2 \leq (1 + A\alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z) - f_i(\check{x})) - \frac{3}{4}\|x_i(k+1) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\|v_i(k) - z\|^2 + B\alpha_k^2 G_f^2.$$

where $A = 8L^2 + 16cL^2$ and $B = 8c + 8$. Here, we have also used Assumption 1(d), according to which the gradients $\nabla f_i(x)$ are bounded on the set $\mathcal{X}$, i.e., $\|\nabla f_i(\Pi_{\mathcal{X}}[v_i(k)])\| \leq G_f$ for all k and i.

Letting $\check{x} = z = \Pi_{\mathcal{X}}[v_i(k)]$ in the preceding relation, we find

$$\|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 \leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X})$$
$$- \frac{3}{4}\|x_i(k+1) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \tag{8}$$

By the definition of the projection, we have

$$\text{dist}(x_i(k+1), \mathcal{X}) = \|x_i(k+1) - \Pi_{\mathcal{X}}[x_i(k+1)]\|$$
$$\leq \|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\|,$$

$$\|x_i(k+1) - v_i(k)\| \geq \left\|\Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k)] - v_i(k)\right\|$$
$$= \text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)}).$$

Upon substituting these estimates in (8), we obtain

$$\text{dist}^2(x_i(k+1), \mathcal{X}) \leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X})$$
$$- \frac{3}{4}\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)})$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \tag{9}$$

Taking the expectation in (9) conditioned on $\mathcal{F}_k$, and using

$$\mathsf{E}\left[\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k\right] \geq \frac{1}{c}\text{dist}^2(v_i(k), \mathcal{X}),$$

which follows by Assumption 3, we find that almost surely

$$\mathsf{E}\left[\text{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k\right] \leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X})$$
$$- \left(\frac{3}{8c} - 2\alpha_k L\right)\text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \tag{10}$$

By using the definition of $v_i(k)$ (as a convex combination of $x_j(k)$ in (3)) and the convexity of the distance function $x \mapsto \mathrm{dist}^2(x, \mathcal{X})$ (see [38, p. 88]), we find that

$$\mathrm{dist}^2(v_i(k), \mathcal{X}) \leq \sum_{j=1}^{m} [W(k)]_{ij} \, \mathrm{dist}^2(x_j(k), \mathcal{X}).$$

The preceding relation and (10) imply that almost surely for all $k \geq 0$,

$$\mathsf{E}\Big[\mathrm{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k\Big]$$

$$\leq (1 + A\alpha_k^2) \sum_{j=1}^{m} [W(k)]_{ij} \, \mathrm{dist}^2(x_j(k), \mathcal{X})$$

$$- \left(\frac{3}{8c} - 2\alpha_k L\right) \mathrm{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2.$$

Finally, by summing over all i and using the fact that each $W(k)$ has column sums equal to 1, we arrive at the following relation: almost surely for all $k \geq 0$,

$$\mathsf{E}\Big[\sum_{i=1}^{m} \mathrm{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k\Big]$$

$$\leq (1 + A\alpha_k^2) \sum_{j=1}^{m} \mathrm{dist}^2(x_j(k), \mathcal{X})$$

$$- \left(\frac{3}{8c} - 2\alpha_k L\right) \sum_{i=1}^{m} \mathrm{dist}^2(v_i(k), \mathcal{X}) + mB\alpha_k^2 G_f^2.$$

Since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, it follows that $\alpha_k \to 0$, implying that there exists $\bar{k}$ such that $\frac{3}{8c} - 2\alpha_k L > 0$ for all $k \geq \bar{k}$. Therefore, for all $k \geq \bar{k}$, all the conditions of the supermartingale theorem are satisfied (Theorem 1). By applying the supermartingale theorem (to a time-delayed process from $\bar{k}$ onward) we conclude that

$$\sum_{k=0}^{\infty} \mathrm{dist}^2(v_i(k), \mathcal{X}) < \infty \qquad \text{for all } i \in V \text{ a.s.}$$

∎

Lemma 5 shows that the points $v_i(k)$ are getting close to the set $\mathcal{X}$ relatively fast, as $k \to \infty$. If the set $\mathcal{X}$ was compact, this would imply that all accumulation points of $\{v_i(k)\}$ would lie in the set $\mathcal{X}$. However, there would be no guarantee that the accumulation points of any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ would be the same. Even worse, Lemma 5 would give no information about optimality of any of the accumulation points. In the next section, we provide a result that helps us claim later on that any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ have the same accumulation points.

C. Disagreement Estimate

We now quantify the agent disagreements in time. We measure the disagreements by using the norm $\|v_i(k) - \bar{v}(k)\|$ of the differences between the estimates $v_i(k)$ generated by different agents according the algorithm (2a)-(2b) and their instantaneous average $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^{m} v_\ell(k)$. The proof of our result relies on a lemma (adopted from [17, Theorem 4.2]), which states that the iterates generated by a "perturbed" consensus protocol are guaranteed to arrive at a consensus when the perturbations are small in some sense. This lemma is provided next.

Lemma 6: Let Assumptions 4 and 5 hold. Consider the iterates generated by

$$\theta_i(k+1) = \sum_{j=1}^{m} [W(k)]_{ij}\theta_j(k) + \epsilon_i(k) \text{ for all } i \in V. \tag{11}$$

Suppose there exists a non-negative non-increasing scalar sequence $\{\alpha_k\}$ such that $\sum_{k=0}^{\infty} \alpha_k \|\epsilon_i(k)\| < \infty$ for all

$i \in V$. Then, for all $i, j \in V$,

$$\sum_{k=0}^{\infty} \alpha_k \|\theta_i(k) - \theta_j(k)\| < \infty.$$

Using Lemma 6, we prove the following disagreement results that will be important in our analysis later.

Lemma 7: Let Assumptions 1, 4 and 5 hold. Also, assume that the stepsize sequence $\{\alpha_k\}$ is non-increasing and such that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Define

$$e_i(k) = x_i(k+1) - v_i(k) \qquad \text{for all } i \in V \text{ and } k \geq 0.$$

Then, we have almost surely

$$\sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty \quad \text{for all } i \in V, \tag{12}$$

$$\sum_{k=0}^{\infty} \alpha_k \|v_i(k) - \bar{v}(k)\| < \infty \quad \text{for all } i \in V, \tag{13}$$

where $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^{m} v_\ell(k)$.

Proof: Define $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$. Consider $\|e_i(k)\|$, for which we can write

$$\|e_i(k)\| \leq \|x_i(k+1) - z_i(k)\| + \|z_i(k) - v_i(k)\|$$
$$= \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_k \nabla f_i(v_i(k))] - z_i(k) \right\|$$
$$+ \|z_i(k) - v_i(k)\|.$$

Since $\mathcal{X} \subseteq \mathcal{X}_i^{\Omega_i(k)}$ and $z_i(k) \in \mathcal{X}$, we have $z_i(k) \in \mathcal{X}_i^{\Omega_i(k)}$. Using the projection theorem (Lemma 1), we obtain

$$\|e_i(k)\|$$
$$\leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\| + \|z_i(k) - v_i(k)\|$$
$$\leq 2\|v_i(k) - z_i(k)\| + \alpha_k \|\nabla f_i(v_i(k))\|$$
$$\leq 2\|v_i(k) - z_i(k)\| + \alpha_k \|\nabla f_i(z_i(k))\|$$
$$+ \alpha_k \|\nabla f_i(v_i(k)) - \nabla f_i(z_i(k))\|$$
$$\leq (2 + \alpha_0 L)\|v_i(k) - z_i(k)\| + \alpha_k G_f, \tag{14}$$

where the last inequality follows by using $\alpha_k \leq \alpha_0$, the Lipschitz gradient property of f_i and the gradient boundedness property (Assumptions 1(c) and 1(d)). Therefore, applying $(a+b)^2 \leq 2a^2 + 2b^2$ in inequality (14), we have for all $i \in V$ and $k \geq 0$,

$$\|e_i(k)\|^2 \leq 2(2 + \alpha_0 L)^2 \|v_i(k) - z_i(k)\|^2 + 2\alpha_k^2 G_f^2. \tag{15}$$

Recall that we defined $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, so we have $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$. In the light of Lemma 5, we also have $\sum_{k=0}^{\infty} \|v_i(k) - z_i(k)\|^2 < \infty$ almost surely. Since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we conclude that

$$\sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty \quad \text{for all } i \in V \text{ a.s.}$$

By applying the inequality $2ab \leq a^2 + b^2$ to each term $\alpha_k \|e_i(k)\|$, we see that for all $i \in V$ almost surely

$$\sum_{k=0}^{\infty} \alpha_k \|e_i(k)\| \leq \frac{1}{2} \sum_{k=0}^{\infty} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty.$$

Now, we note that $x_i(k+1) = v_i(k) + e_i(k)$ with $v_i(k) = \sum_{j=1}^{m}[W(k)]_{ij}x_j(k)$ and the error $e_i(k)$ satisfying $\sum_{k=0}^{\infty}\alpha_k\|e_i(k)\| < \infty$ almost surely. Therefore, by Lemma 6, it follows that

$$\sum_{k=0}^{\infty}\alpha_k\|x_i(k) - x_j(k)\| < \infty \text{ for all } i \text{ and } j \text{ } a.s. \tag{16}$$

Next, we consider $\|v_i(k) - \bar{v}(k)\|$. Recalling that $v_i(k) = \sum_{j=1}^{m}[W(k)]_{ij}\,x_j(k)$ (see (3)) and $W(k)$ is stochastic (Assumption 5), and by using the convexity of the norm, we obtain

$$\|v_i(k) - \bar{v}(k)\| \le \sum_{j=1}^{m} w_{ij}(k)\,\|x_j(k) - \bar{v}(k)\|$$

$$\le \sum_{j=1}^{m}\left\|x_j(k) - \frac{1}{m}\sum_{\ell=1}^{m}x_\ell(k)\right\|,$$

where in the last equality we use $0 \le [W(k)]_{ij} \le 1$ and $\bar{v}(k) = \frac{1}{m}\sum_{\ell=1}^{m}x_\ell(k)$, which holds since $v_i(k) = \sum_{j=1}^{m}[W(k)]_{ij}\,x_j(k)$ and each $W(k)$ is doubly stochastic. Therefore, by using the convexity of the norm again, we see

$$\sum_{j=1}^{m}\left\|x_j(k) - \frac{1}{m}\sum_{\ell=1}^{m}x_\ell(k)\right\| \le \frac{1}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{m}\|x_j(k) - x_\ell(k)\|.$$

We thus have

$$\alpha_k\|v_i(k) - \bar{v}(k)\| \le \frac{\alpha_k}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{m}\|x_j(k) - x_\ell(k)\|,$$

and by using the relation in (16), we conclude that

$$\sum_{k=0}^{\infty}\alpha_k\|v_i(k) - \bar{v}(k)\| < \infty \qquad \text{for all } i \in V \text{ } a.s.$$

∎

V. ALMOST SURE CONVERGENCE OF DRP ALGORITHM

We are now ready to assert the convergence of the method (2a)-(2b) using the lemmas established in Section IV. To outline the rough idea of the proof, let us note that Lemma 5 allows us to infer that $v_i(k)$ approaches the set $\mathcal{X}$. Lemma 7 will allow us to claim that any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ have the same accumulation almost surely, under some mild assumptions on the stepsize. To claim the convergence of the iterates to an optimal solution, it remains to relate the accumulation points of $\{v_i(k)\}$ to the optimal solutions of problem (1). This last piece is provided by the iterate relation of Lemma 4, supported by the supermartingale theorem.

From here onward, we use the following notation regarding the optimal value and optimal solutions of problem (1):

$$f^* = \min_{x\in\mathcal{X}} f(x), \qquad \mathcal{X}^* = \{x \in \mathcal{X} \mid f(x) = f^*\}.$$

We have the following convergence result.

Proposition 1: Let Assumptions 1-5 hold. Let the stepsize be such that $\sum_{k=0}^{\infty}\alpha_k = \infty$ and $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$. Assume that problem (1) has a nonempty optimal set $\mathcal{X}^*$. Then, the iterates $\{x_i(k)\}$, $i \in V$, generated by the method (2a)-(2b) converge almost surely to some random point in the optimal set $\mathcal{X}^*$, i.e., for some random vector $x^\star \in \mathcal{X}^*$,

$$\lim_{k\to\infty} x_i(k) = x^\star \qquad \text{for all } \quad i \in V \text{ } a.s.$$

Proof: We use the definition of the iterate $x_i(k)$ in (2a)-(2b) and lemma 4 with the following identification: $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$, $y = x_i(k+1)$, $x = v_i(k)$, $z = z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, $\alpha = \alpha_k$ and $\tau = c$ where c is the constant from

the relation (4). Thus, for any $\check{x} \in \mathcal{X}$, $k \geq 0$ and $i \in V$, we have

$$\|x_i(k+1) - \check{x}\|^2 \leq (1 + A\alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z_i(k)) - f_i(\check{x})) - \frac{3}{4}\|x_i(k+1) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\|v_i(k) - z_i(k)\|^2 + B\alpha_k^2\|\nabla f_i(\check{x})\|^2,$$

with $A = 8L^2 + 16cL^2$ and $B = 8c + 8$. We next sum the preceding relations over $i = 1, \ldots, m$. Also, we use the convexity of the squared-norm (cf. (5)) and the doubly stochasticity of the weights to obtain the following relation:

$$\sum_{i=1}^m \|v_i(k) - \check{x}\|^2 \leq \sum_{i=1}^m \sum_{j=1}^m [W(k)]_{ij}\|x_j(k) - \check{x}\|^2$$
$$= \sum_{j=1}^m \left(\sum_{i=1}^m [W(k)]_{ij}\right)\|x_j(k) - \check{x}\|^2$$
$$= \sum_{j=1}^m \|x_j(k) - \check{x}\|^2.$$

By doing so, and taking into account that the gradients $\|\nabla f_i(\check{x})\|$ are bounded over $\mathcal{X}$ by a scalar G_f (Assumption 1(d)), we obtain for any $\check{x} \in \mathcal{X}$ and $k \geq 0$,

$$\sum_{i=1}^m \|x_i(k+1) - \check{x}\|^2 \leq (1 + A\alpha_k^2)\sum_{i=1}^m \|x_i(k) - \check{x}\|^2$$
$$- 2\alpha_k \sum_{i=1}^m (f_i(z_i(k)) - f_i(\check{x})) - \frac{3}{4}\sum_{i=1}^m \|x_i(k+1) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 + mB\alpha_k^2 G_f^2. \tag{17}$$

Let $\bar{z}(k) \triangleq \frac{1}{m}\sum_{\ell=1}^m z_\ell(k)$ and recall that $f(x) = \sum_{i=1}^m f_i(x)$. Using $\bar{z}(k)$ and f, we can rewrite the second term on the right hand side in (17) as follows.

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\check{x})) = \sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k)))$$
$$+ (f(\bar{z}(k)) - f(\check{x})). \tag{18}$$

We estimate the first term on the right hand side of the above equation as follows. Using the convexity of each function f_i, we obtain

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \geq \sum_{i=1}^m \langle \nabla f_i(\bar{z}(k)), z_i(k) - \bar{z}(k)\rangle$$
$$\geq -\sum_{i=1}^m \|\nabla f_i(\bar{z}(k))\|\,\|z_i(k) - \bar{z}(k)\|.$$

Since $\bar{z}(k)$ is a convex combination of points $z_i(k) \in \mathcal{X}$, it follows that $\bar{z}(k) \in \mathcal{X}$. This observation and Assumption 1(d), stating that the gradients $\nabla f_i(x)$ are uniformly bounded for $x \in \mathcal{X}$, yield

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \geq -G_f \sum_{i=1}^m \|z_i(k) - \bar{z}(k)\|. \tag{19}$$

We next consider the term $\|z_i(k) - \bar{z}(k)\|$, for which by using $\bar{z}(k) \triangleq \frac{1}{m} \sum_{\ell=1}^{m} z_\ell(k)$ we have

$$\|z_i(k) - \bar{z}(k)\| = \left\| \frac{1}{m} \sum_{\ell=1}^{m} (z_i(k) - z_\ell(k)) \right\|$$

$$\leq \frac{1}{m} \sum_{\ell=1}^{m} \|z_i(k) - z_\ell(k)\| \leq \frac{1}{m} \sum_{\ell=1}^{m} \|v_i(k) - v_\ell(k)\|,$$

where the first inequality is obtained by the convexity of the norm (see (5)) and the last inequality follows by the non-expansive projection property (Lemma 1). Furthermore, by using $\|v_i(k) - v_\ell(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \|v_\ell(k) - \bar{v}(k)\|$, we obtain for every $i \in V$,

$$\|z_i(k) - \bar{z}(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \frac{1}{m} \sum_{\ell=1}^{m} \|v_\ell(k) - \bar{v}(k)\|.$$

Upon summing over $i \in V$, we find that

$$\sum_{i=1}^{m} \|z_i(k) - \bar{z}(k)\| \leq 2 \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\|. \tag{20}$$

Combining relations (20) and (19), and substituting the resulting relation in equation (18), we find that

$$\sum_{i=1}^{m} (f_i(z_i(k)) - f_i(\check{x})) \geq - 2G_f \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\|$$

$$+ (f(\bar{z}(k)) - f(\check{x})).$$

Finally, by using the preceding estimate in inequality (17), we obtain for any $\check{x} \in \mathcal{X}$ and $k \geq 0$,

$$\sum_{i=1}^{m} \|x_i(k+1) - \check{x}\|^2 \leq (1 + A\alpha_k^2) \sum_{i=1}^{m} \|x_i(k) - \check{x}\|^2$$

$$- 2\alpha_k(f(\bar{z}(k)) - f(\check{x})) - \frac{3}{4} \sum_{i=1}^{m} \|x_i(k+1) - v_i(k)\|^2$$

$$+ \left(\frac{3}{8c} + 2\alpha_k L \right) \sum_{i=1}^{m} \|v_i(k) - z_i(k)\|^2$$

$$+ 4\alpha_k G_f \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \tag{21}$$

By the definition of $x_i(k+1)$, we have $x_i(k+1) \in \mathcal{X}_i^{\Omega_i(k)}$, which implies $\|x_i(k+1) - v_i(k)\| \geq \mathrm{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)})$ for $i \in V$. Also, from the definition of $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, we have $\|v_i(k) - z_i(k)\| = \mathrm{dist}(v_i(k), \mathcal{X})$ for $i \in V$. Using these relations and letting $\check{x} = x^*$ for an arbitrary $x^* \in \mathcal{X}^*$, from (21) we obtain for all $k \geq 0$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x^*\|^2 \leq (1 + A\alpha_k^2) \sum_{i=1}^{m} \|x_i(k) - x^*\|^2$$

$$- 2\alpha_k(f(\bar{z}(k)) - f^*) - \frac{3}{4} \sum_{i=1}^{m} \mathrm{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)})$$

$$+ \left(\frac{3}{8c} + 2\alpha_k L \right) \sum_{i=1}^{m} \mathrm{dist}^2(v_i(k), \mathcal{X})$$

$$+ 4\alpha_k G_f \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2.$$

By taking the expectation conditioned on $\mathcal{F}_k$, and noting that $x_i(k)$, $v_i(k)$, $\bar{v}(k)$, and $\bar{z}(k)$ are fully determined

by $\mathcal{F}_k$, we have almost surely for all $x^* \in \mathcal{X}$ and $k \geq 0$,

$$\mathsf{E}\Big[\sum_{i=1}^{m}\|x_i(k+1) - x^*\|^2 \mid \mathcal{F}_k\Big]$$

$$\leq (1 + A\alpha_k^2)\sum_{i=1}^{m}\|x_i(k) - x^*\|^2 - 2\alpha_k(f(\bar{z}(k)) - f^*)$$

$$-\frac{3}{4}\mathsf{E}\Big[\sum_{i=1}^{m}\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k\Big]$$

$$+ \Big(\frac{3}{8c} + 2\alpha_k L\Big)\sum_{i=1}^{m}\text{dist}^2(v_i(k), \mathcal{X})$$

$$+ 4\alpha_k G_f\sum_{i=1}^{m}\|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2.$$

By Assumption 3, we have $\text{dist}^2(x, \mathcal{X}) \leq c\mathsf{E}\Big[\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k\Big]$ for all $x \in \mathcal{X}$ and all $i \in V$. Furthermore, since $\alpha_k \to 0$, by choosing $\bar{k}$ large enough so that $2\alpha_k L \leq \frac{3}{8c}$, we have for all $k \geq \bar{k}$,

$$-\frac{3}{4}\mathsf{E}\Big[\sum_{i=1}^{m}\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k\Big]$$

$$+ \Big(\frac{3}{8c} + 2\alpha_k L\Big)\sum_{i=1}^{m}\text{dist}^2(v_i(k), \mathcal{X}) \leq 0.$$

Thus, we obtain almost surely for all $k \geq \bar{k}$ and $x^* \in \mathcal{X}^*$,

$$\mathsf{E}\Big[\sum_{i=1}^{m}\|x_i(k+1) - x^*\|^2 \mid \mathcal{F}_k\Big]$$

$$\leq (1 + A\alpha_k^2)\sum_{i=1}^{m}\|x_i(k) - x^*\|^2$$

$$- 2\alpha_k(f(\bar{z}(k)) - f^*)$$

$$+ 4\alpha_k G_f\sum_{i=1}^{m}\|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \tag{22}$$

Since $\bar{z}(k) \in \mathcal{X}$, we have $f(\bar{z}(k)) - f^* \geq 0$. Thus, under the assumption $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$ and Lemma 7, relation (22) satisfies all the conditions of the supermartingale convergence of Theorem 1. Hence, the sequence $\{\|x_i(k) - x^*\|^2\}$ is convergent almost surely for any $i \in V$ and $x^* \in \mathcal{X}^*$, and

$$\sum_{k=0}^{\infty}\alpha_k(f(\bar{z}(k)) - f(x^*)) < \infty \quad a.s.$$

The preceding relation and the condition $\sum_{k=0}^{\infty}\alpha_k = \infty$ imply that

$$\liminf_{k \to \infty}(f(\bar{z}(k)) - f(x^*)) = 0 \quad a.s. \tag{23}$$

By Lemma 5, noting that $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, we have $\sum_{k=1}^{\infty}\sum_{i=1}^{m}\|v_i(k) - z_i(k)\|^2 < \infty$ almost surely, implying

$$\lim_{k \to \infty}\|v_i(k) - z_i(k)\| = 0 \quad \text{for all } i \in V \text{ a.s.} \tag{24}$$

Recall that the sequence $\{\|x_i(k) - x^*\|\}$ is convergent almost surely for all $i \in V$ and every $x^* \in \mathcal{X}^*$. Then, in view of relation (2a), we have that the sequence $\{\|v_i(k) - x^*\|\}$ is also convergent almost surely for all $i \in V$ and $x^* \in \mathcal{X}^*$. By relation (24) it follows that $\{\|z_i(k) - x^*\|\}$ is also convergent almost surely for all $i \in V$ and

$x^* \in \mathcal{X}^*$. Since $\|\bar{v}_i(k) - x^*\| \leq \frac{1}{m}\sum_{i=1}^{m}\|v_i(k) - x^*\|$ and the sequence $\{\|v_i(k) - x^*\|\}$ is convergent almost surely for all $i \in V$ and $x^* \in \mathcal{X}^*$, it follows that $\{\|\bar{v}(k) - x^*\|\}$ is convergent almost surely for all $x^* \in \mathcal{X}^*$. Using a similar argument, we can conclude that $\{\|\bar{z}(k) - x^*\|\}$ is convergent almost surely for all $x^* \in \mathcal{X}^*$. As a particular consequence, it follows that the sequences $\{\bar{v}(k)\}$ and $\{\bar{z}(k)\}$ are almost surely bounded and, hence, they have accumulation points. From relation (23) and the continuity of f, it follows that the sequence $\{\bar{z}(k)\}$ must have one accumulation point in the set $\mathcal{X}^*$ almost surely. This and the fact that $\{\|\bar{z}(k) - x^*\|\}$ is convergent almost surely for every $x^* \in \mathcal{X}^*$ imply that for a random point $x^\star \in \mathcal{X}^*$,

$$\lim_{k \to \infty} \bar{z}(k) = x^\star \quad a.s. \tag{25}$$

Now, from $\bar{z}(k) = \frac{1}{m}\sum_{\ell=1}^{m} z_\ell(k)$ and $\bar{v}(k) = \frac{1}{m}\sum_{i=\ell}^{m} v_\ell(k)$, using relation (24) and the convexity of the norm (cf. (5)), we obtain almost surely

$$\lim_{k \to \infty} \|\bar{v}(k) - \bar{z}(k)\| \leq \frac{1}{m}\sum_{\ell=1}^{m} \lim_{k \to \infty} \|v_\ell(k) - z_\ell(k)\| = 0.$$

In view of relation (25), it follows that

$$\lim_{k \to \infty} \bar{v}(k) = x^* \quad a.s. \tag{26}$$

By relation (13) in Lemma 7, we have

$$\liminf_{k \to \infty} \|v_i(k) - \bar{v}(k)\| = 0 \quad \text{for all } i \in V \ a.s. \tag{27}$$

The fact that $\{\|v_i(k) - x^*\|\}$ is convergent almost surely for all i, together with (26) and (27) implies that

$$\lim_{k \to \infty} \|v_i(k) - x^\star\| = 0 \quad \text{for } i \in V \ a.s. \tag{28}$$

Finally, from relation (12) in Lemma 7, we have $\lim_{k \to \infty} \|x_i(k+1) - v_i(k)\| = 0$ for all $i \in V$ almost surely, which together with the limit in (28) yields $\lim_{k \to \infty} x_i(k) = x^\star$ for all $i \in V$ almost surely. ∎

VI. DISTRIBUTED MINI-BATCH RANDOM PROJECTION ALGORITHM

As an extension of the algorithm in (2a)–(2b), one may consider an algorithm where the agents use several random projections at each iteration. Namely, after generating $v_i(k)$ each agent may take (or nature may reveal them) several random samples $\Omega_i^1(k), \ldots, \Omega_i^b(k)$, where each $\Omega_i^r(k) \in I_i$ and $b \geq 1$ is the batch-size. Each collection $\Omega_i^1(k), \ldots, \Omega_i^b(k)$ consists of mutually independent random variables and is independent of the past realizations. More specifically, we have b random independent samples of the *iid* random variable $\Omega_i(k)$ (taking values in I_i). Using the compact form (3) for the update in (2a), in the mini-batch version of the algorithm, each agent $i \in V$, performs the following steps:

$$v_i(k) = \sum_{j=1}^{m} [W(k)]_{ij}\, x_j(k), \tag{29a}$$

$$\psi_i^0(k) = v_i(k) - \alpha_k \nabla f_i(v_i(k)), \tag{29b}$$

$$\psi_i^r(k) = \Pi_{\mathcal{X}_i^{\Omega_i^r(k)}}[\psi_i^{r-1}(k)] \text{ for } r = 1, \ldots, b, \tag{29c}$$

$$x_i(k+1) = \psi_i^b(k), \tag{29d}$$

where $\alpha_k > 0$ is a stepsize at time k and $x_i(0) \in \mathbb{R}^d$ is an initial estimate of agent i (which can be random). The steps in (29b)–(29d) are the successive (random) projections on the sets $\mathcal{X}^{\Omega_i^1(k)}, \ldots, \mathcal{X}^{\Omega_i^b(k)}$ of the point $v_i(k) - \alpha_k \nabla f_i(v_i(k))$.

The algorithm using mini-batches for random projections is of interest when the set I_i is large, i.e., the number of constraint set components $\mathcal{X}_i^j$, $j \in I_i$, of the set $\mathcal{X}_i = \cap_{j \in I_i}\mathcal{X}_i^j$ is large. In such cases, taking several projection steps is beneficial for reducing the infeasibility of the iterates $x_i(k)$ with respect to the set $\mathcal{X}_i$. More concretely, if each set $\mathcal{X}_i$ is the intersection of about 10^4 simpler sets, then one sample of these sets will render a poor approximation of the true set $\mathcal{X}_i$, whereas 100 samples will provide a better approximation of the set. Let $\check{x}$ be a point in the

feasible set $\mathcal{X}$. If just one sample is considered at each iteration, by the non-expansive projection property (Lemma 1), the distance between the next iterate and a point in $\mathcal{X}$ can be estimated as:

$$\|x_i(k+1) - \check{x}\| = \|\psi_i^1(k) - \check{x}\| \leq \|\psi_i^0(k) - \check{x}\|,$$

whereas if 100 samples are considered for projections,

$$\|x_i(k+1) - \check{x}\| = \|\psi_i^{100}(k) - \check{x}\| \leq \ldots$$
$$\leq \|\psi_i^1(k) - \check{x}\| \leq \|\psi_i^0(k) - \check{x}\|,$$

which may yield a larger infeasibility reduction.

For the algorithm using random mini-batch projections, we have the following convergence result.

Proposition 2: Let Assumptions 1-5 hold, and let the stepsize satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Assume that problem (1) has a nonempty optimal set $\mathcal{X}^*$. Then, the iterates $\{x_i(k)\}$, $i \in V$, produced by the method (29a)-(29d) converge to some random point in the optimal set $\mathcal{X}^*$ almost surely, i.e., for some random vector $x^* \in \mathcal{X}^*$,

$$\lim_{k\to\infty} x_i(k) = x^* \qquad \text{for all} \quad i \in V \text{ a.s.}$$

Proof: The proof of this result is similar to that of Proposition 1. It requires some adjustments of Lemma 5 and Lemma 7. The proof with these adjustments is provided in Appendix A. ∎

VII. APPLICATION - DISTRIBUTED SUPPORT VECTOR MACHINES (DRSVM)

In this section, we apply our DRP algorithm and its mini-batch variant to Support Vector Machines (SVMs). We provide a brief introduction to SVMs in Subsection VII-A, while in Subsection VII-B we report our numerical results on some data sets that are generously made available by Thorsten Joachims.

A. Support Vector Machines

Support Vector Machines (SVMs) are popular classification tools with a strong theoretical background. Given a set of n example-label pairs $\{(a_j, b_j)\}_{j=1}^n$, $a_j \in \mathbb{R}^d$ and $b_j \in \{+1, -1\}$, we need to find a vector $x = [y^T \; \boldsymbol{\xi}^T]^T \in \mathbb{R}^{d+n}$ that solves the following optimization problem (a bias term is included in y for convenience):

$$\min_{y,\boldsymbol{\xi}} f(y, \boldsymbol{\xi}) = \frac{1}{2}\|y\|^2 + C \sum_{j=1}^{n} \xi_j \tag{30}$$

$$\text{s.t. } b_j \langle y, a_j \rangle \geq 1 - \xi_j, \; \xi_j \geq 0, \; \text{for all } j \in \{1, \ldots, n\}.$$

Here, we use slack variables ξ_j, for $j = 1, \ldots, n$, to consider linearly non-separable cases as well. If the optimal solution $(y^*, \boldsymbol{\xi}^*)$ to this problem exists, the solution y^* is the maximum-margin separating hyperplane [40].

For applying DRP to problem (30), we can define f_i and $\mathcal{X}_i$, as follows:

$$f_i(x) = \frac{1}{2m}\|y\|^2 + C \sum_{j \in I_i} \xi_j,$$

$$\mathcal{X}_i = \{x \in \mathbb{R}^{d+n} \mid b_j \langle y, a_j \rangle \geq 1 - \xi_j, \; \xi_j \geq 0, \; \forall j \in I_i\}.$$

where I_i is a set of indices such that $\cup_{i=1}^m I_i = \{1, \ldots, n\}$, $I_i \cap I_j = \emptyset$ for $i \neq j$ and $j \in I_i$ if and only if $\mathcal{X}_i$ contains inequalities associated with the data (a_j, b_j). Note that each set $\mathcal{X}_i^j = \{x \in \mathbb{R}^{d+n} \mid b_j \langle y, a_j \rangle \geq 1 - \xi_j, \; \xi_j \geq 0\}$ is the intersection of two halfspaces, the projection onto which can be computed in a few steps (see Appendix B).

B. Simulations

In the section, we perform some experiments with our DRP algorithm. We refer to our DRP algorithm applied on SVMs as DrSVM. The purpose of the experiments is to verify the convergence and to show in how many iterations the proposed method can actually arrive at consensus in distributed settings. We use the DRP algorithm in (2a)-(2b) and its variant in (29a)-(29d) with the stepsize $\alpha_k = \frac{1}{k+1}$ for $k \geq 0$. We vary the number of batches b as 1, 100 or 1000 to observe the different convergence speed, where $b = 1$ corresponds to the algorithm in (2a)-(2b). To show

TABLE I
THE STATISTICS OF THREE TEXT CLASSIFICATION DATA SETS: n IS THE NUMBER OF EXAMPLES AND d IS THE NUMBER OF FEATURES. s REPRESENTS THE SPARSITY OF DATA.

Data set	Statistics		
	n	d	s
astro-ph	62,369	99,757	0.08%
CCAT	804,414	47,236	0.16%
C11	804,414	47,236	0.16%

the effect of connectivity, we compare two different time invariant network topologies, i) a completely connected graph (clique) and ii) a 3-regular expander graph. The 3-regular expander graph is a sparse graph that has strong connectivity with every node having degree 3.

We use 3 text classification data sets for our experiments. The data sets were kindly provided by Thorsten Joachims (see [5] for their descriptions). Table I lists the statistics of the data sets. All of the data sets are from binary document classification. Since the data sets used here are very unlikely separable, we use the formulation (30) with $C = 1$. In each experiment the number n of constraints is divided among the agents equally (if n is not divisible by m, the m-th agent gets the remainder). To estimate the generalization (or testing) performance, we split the data and use 80% for training and 20% for testing.

DrSVM is implemented with C/C++ and all experiments were performed on a 64-bit machine running Fedora 16 with an Intel Core 2 Quad Processor Q9400 and 8G of RAM. The experiments are not performed on a real networked environment so we do not consider delays and node/link failures that may exist in networks.

For stopping criteria, we first run a centralized random incremental projection [31] on the 80% training set with $b = 1$ until the relative error of objective values in two consecutive iterations is less than 0.001. i.e.,

$$|f(x(k)) - f(x(k+1))|/f(x(k)) < 0.001.$$

We then measure the test accuracy of the final solution on the remaining 20% test set, which will become the target test accuracy t_{acc}. For experiments in the distributed setting, we measure the test accuracy of every agent's solution at the end of every iteration. If every solution at certain iteration satisfies the target value t_{acc}, we conclude that the agents arrived at a consensus and the algorithm converged. The maximum number of iterations in each simulation is limited to 20,000.

Table II shows the results. As we do more projections per iteration, the total number of iterations required for convergence is less, regardless of the number of agents. For the given stopping criteria, it seems that less iterations are needed for DrSVM to converge as the number of agents increases. We can also observe the effect of network connectivity. When all the other parameters (m and b) are the same, for most of the cases, the number of iterations required for the 3-regular expander graph to converge is greater or equal to that for the clique.

The table reports the number of iterations required for all the agents to achieve the target test accuracy. Therefore, the total number of projections is *at most* the number of iterations times m times b. This is because no projection is required if the current estimate is already in the selected constraint component. For example, the total number of projections for astro-ph with $m = 6$ and $b = 100$ is at most $4,800(= 8 \times 6 \times 100)$.

The runtime (or the number of calculations) of the algorithm is not only proportional to the number of projections, but also to the number of gradient updates. For example, for astro-ph with $m = 6$ and $b = 1$, the total number of projections is $4,170(= 695 \times 6 \times 1)$, while the total number of gradient updates is $4,170(= 695 \times 6)$. For the same example with $m = 6$ and $b = 100$, the total number of projections is $4,800(= 8 \times 6 \times 100)$, but the total number of gradient updates is only $48(= 8 \times 6)$. In any case, the numbers are much smaller than the number 62,369 of the training data points. This shows that DrSVM can quickly find a good quality solution before examining the training samples even once.

To show the convergence (and consensus) of the algorithm, we plot in Figure 1 the objective value $f(x)$ of centralized random projection (CRP) and DRP with 10 agents for example astro-ph. Note that we plot the convergence of the objective value instead of the solution. This is because CRP and DRP may converge to different optimal points as the problem (30) may not have a unique optimal solution. For Figure 1(a) and 1(b), we applied the random projection once and 100 times per iteration, respectively. From the figures, we can observe that the
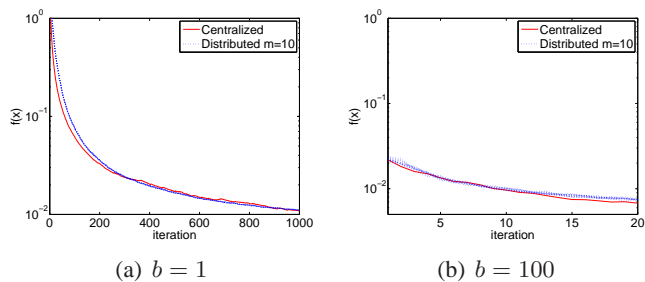
(a) $b = 1$

(b) $b = 100$

Fig. 1. $f(x)$ vs iteration on `astro-ph` with 10 agents when batch size b is 1 and 100.

TABLE II

THE RESULTS OF DRSVM WITH TWO DIFFERENT GRAPH TOPOLOGIES (CLIQUE AND 3-REGULAR EXPANDER GRAPH) AND THREE DIFFERENT NUMBERS OF AGENTS ($m = 2, 6, 10$): t_{acc} IS THE TARGET TEST ACCURACY AND b IS THE NUMBER OF PROJECTIONS PER ITERATION. THE TABLE SHOWS THE NUMBER OF ITERATIONS FOR ALL AGENTS TO REACH THE TARGET TEST ACCURACY, WHERE '-' INDICATES THAT THE ALGORITHM DID NOT CONVERGE WITHIN THE 20,000 MAXIMUM ITERATION LIMIT.

Data set	t_{acc}	b	$m = 2$	Clique		3-regular expander	
				$m = 6$	$m = 10$	$m = 6$	$m = 10$
astro-ph	0.95	1	1,055	695	697	695	-
		100	11	8	11	11	11
		1000	2	2	2	2	2
CCAT	0.91	1	752	511	362	517	-
		100	11	10	8	10	8
		1000	2	3	2	3	3
C11	0.97	1	1,511	1,255	799	1,226	-
		100	16	17	12	17	15
		1000	2	2	2	2	2

objective values of CRP and the 10 agents in DRP are almost identical. The final objective of Figure 1(b) seems smaller than that of 1(a). This is because the stepsize at iteration 1000 is too small.

VIII. CONCLUSIONS

We have proposed and analyzed a distributed gradient algorithm with random incremental projections for a network of agents with time-varying connectivity. We considered the most general cases, where each agent has a unique and different objective and constraint. The proposed algorithm is applicable to problems where the whole constraint set is not known in advance but its component is revealed in time, or where the projection onto the whole set is computationally prohibitive. We have established almost sure convergence of the algorithm when the objective is convex under typical assumptions. Also, we have provided a variant of the algorithm using a mini-batch of consecutive projections and established its convergence in almost sure sense. Experiments on three text classification benchmarks using SVMs were performed to verify the performance of the proposed algorithm.

Future work includes some extensions of the distributed model proposed here. First, we have assumed the gradients evaluated have no errors. We can consider the effects of stochastic gradient errors in the future analysis. Second, more robust algorithms can be developed to also handle asynchronous networks with communication delays, noise and/or failures in links/nodes. Third, an implementation of a real parallel computing environment will be needed to handle large-scale data sets.

APPENDIX

A. Proof of Proposition 2

We construct the proof by adjusting the result of Lemma 4, and by verifying that Lemma 5 and Lemma 7 apply to the mini-batch variant of the DRP method. The basic insight that guides the proof is that the operation of

successive projections on components $\mathcal{X}_i^j$ of the set $\mathcal{X}_i = \cap_{j \in I_i} \mathcal{X}_j^j$ remains a non-expansive operation with respect to points that belong to the set $\mathcal{X}_i$, as well as with respect to the points in the intersection set $\mathcal{X} = \cap_{i=1}^m \mathcal{X}_i$.

1) Basic Iterate Relation for Mini-Batch Algorithm: For the iterates generated by the mini-batch random projection algorithm in (29a)–(29d), we have the following basic result.

Lemma 8: Let Assumption 1 hold. Then, for any $\check{x} \in \mathcal{X}$, and for all $i \in V$ and all $k \geq 0$,

$$\|x_i(k+1) - \check{x}\|^2 \leq (1 + A_\tau \alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z) - f_i(\check{x})) - \frac{3}{4}\|\psi_i^1(k) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8\tau} + 2\alpha_k L\right)\|v_i(k) - z\|^2 + B_\tau \alpha_k^2 G_f^2,$$

where $A_\tau = 8L^2 + 16\tau L^2$, $B_\tau = 8\tau + 8$ and $\tau > 0$ is arbitrary.

Proof: By using the non-expansiveness property of projection operation (Lemma 1(a)), we have for arbitrary $\check{x} \in \mathcal{X}$ (since $\mathcal{X} \subseteq \mathcal{X}_i^j$ for all $j \in I_i$), and for all $i \in V$ and $k \geq 0$,

$$\|x_i(k+1) - \check{x}\| \leq \|\psi_i^{b-1}(k) - \check{x}\|$$
$$\leq \cdots \leq \|\psi_i^1(k) - \check{x}\|. \tag{31}$$

The intermediate iterate $\psi_i^1(k)$ is just obtained after one projection step,

$$\psi_i^1(k) = \Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k) - \alpha_k \nabla f_i(v_i(k))],$$

so it satisfies Lemma 4 with $y = \psi_i^1(k)$, $\mathcal{Y} = \mathcal{X}_i^{\Omega_i^1(k)}$, $x = v_i(k)$, $\alpha = \alpha_k$, and $\phi = f_i$. Thus, we have for any $\check{x} \in \mathcal{X}$ and $z \in \mathbb{R}^d$,

$$\|\psi_i^1(k) - \check{x}\|^2 \leq (1 + A_\tau \alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z) - f_i(\check{x})) - \frac{3}{4}\|\psi_i^1(k) - v_i(k)\|^2$$
$$+ \left(\frac{3}{8\tau} + 2\alpha_k L\right)\|v_i(k) - z\|^2 + B_\tau \alpha_k^2 \|\nabla f_i(\check{x})\|^2. \tag{32}$$

From (31) and (32), by using the gradient boundedness property of Assumption 1(d), we obtain the stated relation. ∎

2) Conditional Expectation Relation for Mini-Batch Algorithm: The convergence proof of Proposition 2 requires a relation for the iterates involving expectations with respect to the past history of the method. For this, we need to define a relevant σ-algebra. We let $\tilde{\mathcal{F}}_k$ be the σ-algebra generated by the entire history of the algorithm up to time $k-1$ inclusively. Thus, $\tilde{\mathcal{F}}_k$ includes the realizations of all the random variables but not the realizations of the indices $\Omega_i^1(k), \ldots, \Omega_i^b(k)$ at time k. Specifically, it is given by for all $k \geq 1$,

$$\tilde{\mathcal{F}}_k = \{x_i(0), i \in V\}$$
$$\cup \{\Omega_i^r(\ell); 0 \leq \ell \leq k-1, 1 \leq r \leq b, i \in V\}$$

where $\tilde{\mathcal{F}}_0 = \{x_i(0), i \in V\}$.

Now, with this definition of the σ-algebra, we have the following result.

Lemma 9: Let Assumptions 1 and 3 hold. Then, almost surely for any $\check{x} \in \mathcal{X}$, and for all $i \in V$ and all $k \geq 0$,

$$\mathsf{E}\left[\|x_i(k+1) - \check{x}\|^2 \mid \tilde{\mathcal{F}}_k\right] \leq (1 + A\alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z_i(k)) - f_i(\check{x}))$$
$$- \left(\frac{3}{8c} - 2\alpha_k L\right)\text{dist}^2(v_i x(k), \mathcal{X}) + B\alpha_k^2 G_f^2.$$

where $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, $A = 8L^2 + 16cL^2$, $B = 8c + 8$, and c is from Assumption 3.

Proof: By letting $z = z_i(k)$ and $\tau = c$ in Lemma 8, we obtain

$$\mathsf{E}\Big[\|x_i(k+1) - \check{x}\|^2 \mid \tilde{\mathcal{F}}_k\Big] \leq (1 + A\alpha_k^2)\|v_i(k) - \check{x}\|^2$$
$$- 2\alpha_k(f_i(z_i(k)) - f_i(\check{x})) - \frac{3}{4}\mathsf{E}\Big[\|\psi_i^1(k) - v_i(k)\|^2 \mid \tilde{\mathcal{F}}_k\Big]$$
$$+ \left(\frac{3}{8c} + 2\alpha_k L\right)\|v_i(k) - z_i(k)\|^2 + B\alpha_k^2 G_f^2,$$

where $A = 8L^2 + 16cL^2$ and $B = 8c + 8$.

Since $\psi_i^1(k) \in \mathcal{X}_i^{\Omega_i^1(k)}$, by the projection property we have $\|\psi_i^1(k) - v_i(k)\|^2 \geq \|\Pi_{\mathcal{X}^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2$. Then,

$$\mathsf{E}\Big[\|\psi_i^1(k) - v_i(k)\|^2 \tilde{\mathcal{F}}_k\Big]$$
$$\geq \mathsf{E}\Big[\|\Pi_{\mathcal{X}^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid \tilde{\mathcal{F}}_k\Big]$$
$$= \mathsf{E}\Big[\|\Pi_{\mathcal{X}^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid v_i(k)\Big].$$

Furthermore, by Assumption 3 we have

$$\mathsf{E}\Big[\|\Pi_{\mathcal{X}^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid v_i(k)\Big] \geq \frac{1}{c}\mathrm{dist}^2(v_i(k), \mathcal{X}).$$

The preceding relations and $\mathrm{dist}(v_i(k), \mathcal{X}) = \|v_i(k) - z_i(k)\|$ yield the desired relation. ∎

3) Lemma 5 and Lemma 7 hold: Using Lemma 9, we argue that the results of Lemma 5 and Lemma 7 apply to the mini-batch random projection algorithm.

Claim 1: Lemma 5 holds for the iterates generated by method (29a)–(29d).

Proof: By letting $\check{x} = \Pi_{\mathcal{X}}[v_i(k)]$ in Lemma 9, and noting that $\|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\| \geq \mathrm{dist}(x_i(k+1), \mathcal{X})$ and $\|v_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\| = \mathrm{dist}(v_i(k), \mathcal{X})$, we obtain

$$\mathsf{E}\Big[\mathrm{dist}^2(x_i(k+1), \mathcal{X}) \mid \tilde{\mathcal{F}}_k\Big] \leq (1 + A\alpha^2)\mathrm{dist}^2(v_i(k), \mathcal{X})$$
$$- \left(\frac{3}{8c} - 2\alpha_k L\right)\mathrm{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2,$$

which is the same as relation (10) within the proof of Lemma 5. The rest of the proof of Lemma 5 holds exactly as given, and the result of Lemma 5 remains valid. ∎

Claim 2: Lemma 7 holds for the iterates generated by method (29a)–(29d).

Proof: Define $e_i(k) = x_i(k+1) - v_i(k)$ and $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$. Now, consider $\|e_i(k)\|$ for which we have

$$\|e_i(k)\| \leq \|x_i(k+1) - z_i(k)\| + \|z_i(k) - v_i(k)\|.$$

The non-expansiveness projection property and the fact $z_i(k) \in \mathcal{X} \subset \mathcal{X}_i^{\Omega_i^r(k)}$, for all $r = 1, \ldots, b$, and any realization of these sets imply

$$\|x_i(k+1) - z_i(k)\|$$
$$\leq \|\psi_i^{b-1}(k) - z_i(k)\| \leq \cdots \leq \|\psi_i^1(k) - z_i(k)\|$$
$$\leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\|.$$

Therefore

$$\|e_i(k)\| \leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\| + \|z_i(k) - v_i(k)\|,$$

which is the same as the first inequality in (14) within the proof of Lemma 7. The rest of the proof of that lemma holds in verbatim, and the result follows. ∎

4) Details of the Proof of Proposition 2: We now connect the preceding results and provide the proof of Proposition 2. Starting from the relation in Lemma 9, after summing over all $i \in V$, we can see that almost surely

for all $\check{x} \in \mathcal{X}$ and all $k \geq 0$,

$$\mathsf{E}\Big[\sum_{i=1}^{m} \|x_i(k+1) - \check{x}\|^2 \mid \tilde{\mathcal{F}}_k\Big]$$

$$\leq (1 + A\alpha_k^2) \sum_{i=1}^{m} \|v_i(k) - \check{x}\|^2$$

$$- 2\alpha_k \sum_{i=1}^{m} (f_i(z_i(k)) - f_i(\check{x}))$$

$$- \Big(\frac{3}{8c} - 2\alpha_k L\Big) \sum_{i=1}^{m} \|v_i(k) - z_i(k)\|^2 + mB\alpha_k^2 G_f^2,$$

where $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$.

Now, the same as in the proof of Proposition 1, using the properties of the matrices $W(k)$ and the convexity of the squared-norm function (see (5)), we can show that

$$\sum_{i=1}^{m} \|v_i(k) - \check{x}\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - \check{x}\|^2.$$

Also, using verbatim arguments, we can show that

$$\sum_{i=1}^{m} (f_i(z_i(k)) - f_i(\check{x}))$$

$$\geq -2G_f \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\| + (f(\bar{z}(k) - f(\check{x})),$$

where $\bar{z}(k) = \frac{1}{m} \sum_{\ell=1}^{m} z_\ell(k)$ and $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^{m} v_\ell(k)$. Under the conditions of Proposition 2, we have $\alpha_k \to 0$. Choosing $\bar{k}$ large enough so that $2\alpha_k L \leq \frac{3}{8c}$ for all $k \geq \bar{k}$, we have

$$- \Big(\frac{3}{8c} - 2\alpha_k L\Big) \sum_{i=1}^{m} \|v_i(k) - z_i(k)\|^2 \leq 0.$$

By combining all the preceding relations, we obtain almost surely for all $\check{x} \in \mathcal{X}$ and all $k \geq \bar{k}$,

$$\mathsf{E}\Big[\sum_{i=1}^{m} \|x_i(k+1) - \check{x}\|^2 \mid \tilde{\mathcal{F}}_k\Big]$$

$$\leq (1 + A\alpha_k^2) \sum_{i=1}^{m} \|x_i(k) - \check{x}\|^2$$

$$- 2\alpha_k (f(\bar{z}(k)) - f(\check{x}))$$

$$+ 4\alpha_k G_f \sum_{i=1}^{m} \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2.$$

Letting $\check{x} = x^*$ for an arbitrary optimal solution $x^* \in \mathcal{X}^*$, from the preceding relation we arrive at relation (22) in the proof of Proposition 1. From relation (22) onward, the proof of Proposition 1 holds verbatim, and the stated almost sure convergence of the mini-batch method follows.

B. Projection onto the Intersection of Two Half-spaces

Given $v \in \mathbb{R}^d$, we are interested in solving the following optimization problem.

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - v\|^2 \tag{33}$$

$$\text{s.t.} \quad \langle a, w \rangle \leq b, \ w_i \geq 0,$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ and w_i is the i-th component of the vector w.

The two half-spaces divide the $\mathbb{R}^d$ space into four parts. Therefore, there are only four cases to consider.

1) $\langle a, v \rangle \leq b$ and $v_i \geq 0$.

In this case, v is already in the intersection and $w = v$.

2) $\langle a, v \rangle > b$ and $v_i < 0$.

In this case, v is projected onto the intersection of the two hyperplanes $\{w \mid \langle a, w \rangle = b\}$ and $\{w \mid w_i = 0\}$. Finding such a projection is equivalent to solving the following optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w - v\|^2 \tag{34}$$
$$\text{s.t.} \quad \langle a, w \rangle = b, \ w_i = 0.$$

The Lagrangian of the problem (34) is

$$\mathcal{L}(w, \theta, \zeta) = \frac{1}{2}\|w - v\|^2 + \theta \left(\sum_{j=1}^{d} a_j w_j - b \right) + \zeta w_i,$$

where $\theta, \zeta \in \mathbb{R}$ are Lagrange multipliers. Differentiating the Lagrangian and setting it to zero gives the optimality condition,

$$w_i^* - v_i + a_i \theta^* + \zeta^* = 0,$$
$$w_j^* - v_j + a_j \theta^* = 0, \quad \text{for } j \neq i.$$

From the primal feasibility, we have the following relations:

$$w_i^* = 0 \implies \zeta^* = v_i - a_i \theta^*,$$

$$\sum_{j=1}^{n} a_j w_j^* = \sum_{j \neq i} a_j w_j^* = \sum_{j \neq i} a_j (v_j - a_j \theta^*) = b$$

$$\implies \quad \theta^* = \frac{\sum_{j \neq i} a_j v_j - b}{\sum_{j \neq i} a_j^2}.$$

Therefore, the projection is given by

$$w_j^* = \begin{cases} 0 & \text{if } j = i, \\ v_j - a_j \theta^* & \text{otherwise.} \end{cases}$$

Let $w^* = [w_1^*, \ldots, w_d^*]^T$.

3) $\langle a, v \rangle > b$ and $v_i \geq 0$.

In this case, v will be projected either onto the hyperplane $\{w \mid \langle a, w \rangle = b\}$ or onto the intersection of the two hyperplanes $\{w \mid \langle a, w \rangle = b\}$ and $\{w \mid w_i = 0\}$. Let $\hat{w}$ be the projection of v onto $\{w \mid \langle a, w \rangle = b\}$, i.e.,

$$\hat{w} = v - \left(\frac{\langle a, v \rangle - b}{\|a\|^2} \right) a.$$

The projection of v in this case is given by

$$w = \begin{cases} \hat{w} & \text{if } \hat{w}_i \geq 0, \\ w^* & \text{otherwise.} \end{cases}$$

4) $\langle a, v \rangle \leq b$ and $v_i < 0$.

Let $\hat{w}$ be the projection of v onto the hyperplane $\{w \mid w_i = 0\}$, i.e.,

$$\hat{w} = v - (v_i - b) e_i,$$

where $e_i \in \mathbb{R}^d$ is the vector whose i-th component is one and all the other components are zero. Then, the

projection of v is given by

$$w = \begin{cases} \hat{w} & \text{if } \langle a, \hat{w} \rangle \leq b, \\ w^* & \text{otherwise.} \end{cases}$$

REFERENCES

[1] M. Rabbat and R. D. Nowak, "Distributed optimization in sensor networks," in *IPSN*, 2004, pp. 20–27.

[2] S. S. Ram, V. V. Veeravalli, and A. Nedić, "Distributed non-autonomous power control through distributed convex optimization," in *IEEE INFOCOM*, 2009, pp. 3001–3005.

[3] B. Johansson, "On distributed optimization in networked systems," Ph.D. dissertation, Royal Institute of Technology (KTH), tRITA-EE 2008:065, 2008.

[4] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383 –1400, March 2010.

[5] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 217–226.

[6] B. Johansson, M. Rabi, and M. Johansson, "A simple peer-to-peer algorithm for distributed optimization in sensor networks," in *Proceedings of the 46th IEEE Conference on Decision and Control*, Dec. 2007, pp. 4705 –4710.

[7] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM J. on Optimization*, vol. 20, no. 2, pp. 691–717, Jun. 2009.

[8] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[9] F. Cattivelli and A. H. Sayed, "Diffusion lms strategies for distributed estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1035–1048, March 2010.

[10] C. Lopes and A. H. Sayed, "Distributed processing over adaptive networks," in *Adaptive Sensor Array Processing Workshop, MIT Lincoln Laboratory, MA*, 2006, pp. 1–5.

[11] A. Nedić and A. Ozdaglar, "On the rate of convergence of distributed asynchronous subgradient methods for multi-agent optimization," in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, USA, 2007, pp. 4711–4716.

[12] A. H. Sayed and C. G. Lopes, "Adaptive processing over distributed networks," *IEICE Transactions on Fundamentals of Electronics,Communications and Computer Sciences*, vol. E90-A, no. 8, pp. 1504–1510, 2007.

[13] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48 –61, Jan. 2009.

[14] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, "Distributed subgradient methods and quantization effects," in *Proceedings of 47th IEEE Conference on Decision and Control*, December 2008, pp. 4177 –4184.

[15] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1291 –1306, June 2011.

[16] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010.

[17] ——, "A new class of distributed optimization algorithms: application to regression of distributed data," *Optimization Methods and Software*, vol. 27, no. 1, pp. 71–88, 2012.

[18] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592 –606, March 2012.

[19] A. Nedić, A. Ozdaglar, and P. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922 –938, April 2010.

[20] K. Srivastava, A. Nedić and, and D. Stipanović and, "Distributed constrained optimization over noisy networks," in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, Dec. 2010, pp. 1945 –1950.

[21] K. Srivastava and A. Nedić, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 772–790, 2011.

[22] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[23] A. H. Sayed, "Diffusion adaptation over networks," to appear in E-Reference Signal Processing, R. Chellapa and S. Theodoridis, editors, Elsevier, 2013. Also available online as arXiv:1205.4220v1, May 2012.

[24] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.

[25] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," 2012, http://arxiv.org/abs/1205.3993.

[26] ——, "On the influence of informed agents on learning and adaptation over networks," 2012, http://arxiv.org/abs/1203.1524.

[27] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988 – 1001, June 2003.

[28] R. Olfati-Saber and R. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520 – 1533, Sep. 2004.

[29] L. Xiao, S. P. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.

[30] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 33–55, Feb. 2009.

[31] A. Nedić, "Random algorithms for convex minimization problems," *Mathematical Programming - B*, vol. 129, pp. 225–253, 2011.

[32] B. Polyak, "Random algorithms for solving convex inequalities," in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, ser. Studies in Computational Mathematics, Y. C. Dan Butnariu and S. Reich, Eds.   Elsevier, 2001, vol. 8, pp. 409 – 422.

[33] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, 1984.

[34] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803 – 812, Sep. 1986.

[35] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods.*   Athena Scientific, Belmont, 1997.

[36] J. V. Burke and M. C. Ferris, "Weak sharp minima in mathematical programming," *SIAM Journal on Control and Optimization*, vol. 31, pp. 1340–1359, 1993.

[37] L. Gubin, B. Polyak, and E. Raik, "The method of projections for finding the common point of convex sets," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1 – 24, 1967.

[38] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*.   Athena Scientific, 2003.

[39] B. Polyak, *Intro. to optimization*.   Optimization software, Inc., Publications division, New York, 1987.

[40] V. N. Vapnik, *The nature of statistical learning theory*.   Springer-Verlag New York, Inc., 1995.